

(19)日本国特許庁 (J P)

(12) 公 開 特 許 公 報 (A)

(11)特許出願公開番号

特開平10-257436

(43)公開日 平成10年(1998)9月25日

(51)Int.Cl.⁶
H 0 4 N 5/93
5/92
7/32

識別記号

F I
H 0 4 N 5/93 Z
5/92 H
7/137 Z

審査請求 未請求 請求項の数7 O L (全 14 頁)

(21)出願番号 特願平9-55340

(22)出願日 平成9年(1997)3月10日

(71)出願人 391023987
松下 温
東京都新宿区喜久井町36

(71)出願人 392008231
岡田 謙一
東京都文京区本郷4-25-12

(72)発明者 松下 温
東京都新宿区喜久井町36

(72)発明者 岡田 謙一
東京都文京区本郷4-25-12

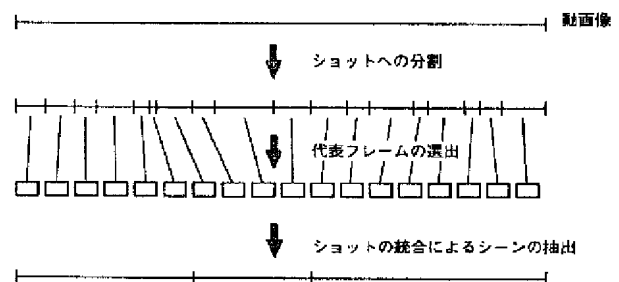
(74)代理人 弁理士 鈴木 正次

(54)【発明の名称】 動画像の自動階層構造化方法及びこれを用いたブラウジング方法

(57)【要約】

【課題】 この発明は、符号化した動画像を自動階層構造化し、この動画像とその解析データを基にしてビデオブラウザを得ることを目的としたものである。

【解決手段】 動画像を符号化し、該符号化された動画像を各ショットに分割し、ついで分割されたショット毎の類似度を用い、ショットを統合してシーンを抽出処理することを特徴とした動画像の自動階層構造化方法。動画像を符号化し、該符号化された動画像を各ショットに分割し、ついで分割されたショット毎の類似度を用い、ショットを統合してシーンを抽出処理して動画像を自動階層構造化し、この階層構造化されたデータを用いて動画像全体の内容把握、所望のシーンまたはショットの検出を容易にすることを特徴とした動画像のブラウジング方法。



【特許請求の範囲】

【請求項1】 符号化された動画を各ショットに分割し、ついで分割されたショット毎の類似度を用い、ショットを統合してシーンを抽出処理することを特徴とした動画の自動階層構造化方法。

【請求項2】 動画の符号化は、MPEGによるものとすることを特徴とした請求項1記載の動画の自動階層構造化方法。

【請求項3】 符号化された動画からショットを検出する際に、MPEGの特徴を利用して高速に処理することを特徴とした請求項1記載の動画の自動階層構造化方法。

【請求項4】 ショット同士の類似度の算出に際し、代表フレームを抽出することを特徴とした請求項1記載の動画の自動階層構造化方法。

【請求項5】 ショット間の類似度をファジィ推論により求めることを特徴とした請求項1記載の動画の自動階層構造化方法。

【請求項6】 シーンの抽出処理は、定義されたショット間の結合度により求めることを特徴とした請求項1記載の動画の自動階層構造化方法。

【請求項7】 符号化された動画を各ショットに分割し、ついで分割されたショット毎の類似度を用い、ショットを統合してシーンを抽出処理して動画を自動階層構造化し、この階層構造化されたデータを用いて動画全体の内容把握、所望のシーンまたはショットの検出を容易にすることを特徴とした動画のブラウジング方法。

【発明の詳細な説明】**【0001】**

【発明の属する技術分野】この発明は、符号化した動画を自動階層構造化し、この動画とその解析データを基にしてビデオブラウザを得ることを目的とした動画の自動階層構造化方法及びこれを用いたブラウジング方法に関する。

【0002】

【従来の技術】現在動画情報は、ビデオの単純再生の域を脱していない。即ち動画はフレーム単位でとらえられており、撮影時に符号を付した場合には、その符号によって抽出できると共に、再生時間をファクターとして所望のフレームを検出又は再生するなど特別な関係が明らかな場合に限り、当該フレームを抽出することができる。

【0003】

【発明により解決すべき課題】然し乍ら何等の符号を付することなく、与えられる動画情報から所望のフレームを抽出することは極めて困難であり、時間的制約があれば、抽出不可能となる。例えば一般に1フレームは30分の1秒であるから、1分間に1800フレーム、1時間に108000フレームとなる。

【0004】そこで前記従来の一般動画情報から任意のフレームを短時間で抽出することができない問題点があった。

【0005】

【課題を解決する為の手段】然るにこの発明は、符号化された動画を各ショットに分割し、ショット毎の類似度を用いてショットを統合し、シーンを抽出することにより自動階層構造化し、このデータを用いて動画のブラウジングツールを作成することによって前記従来の問題点を解決したのである。

【0006】即ちこの発明は、符号化された動画を各ショットに分割し、ついで分割されたショット毎の類似度を用い、ショットを統合してシーンを抽出処理することを特徴とした動画の自動階層構造化方法であり、動画の符号化は、MPEGによるものとすることを特徴としたものである。また符号化された動画からショットを検出する際に、MPEGの特徴を利用して高速に処理することを特徴としたものであり、ショット同士の類似度の算出に際し、代表フレームを抽出することを特徴としたものである。次にショット間の類似度をファジィ推論により求めることを特徴としたものであり、シーンの抽出処理は、定義されたショット間の結合度により求めることを特徴としたものである。更に他の発明は、符号化された動画を各ショットに分割し、ついで分割されたショット毎の類似度を用い、ショットを統合してシーンを抽出処理して動画を自動階層構造化し、この階層構造化されたデータを用いて動画全体の内容把握、所望のシーンまたはショットの検出を容易にすることを特徴とした動画のブラウジング方法である。

【0007】前記における符号化はMPEG1の圧縮アルゴリズムによる。ここにMPEG1の正式名称は「Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s」である。

【0008】前記ハイブリッド符号化は、DCTと量子化、動き補償フレーム間予測及びエントロピー符号化により行うが、前記個々の方法は公知の方法であるから詳細な説明は省略する。

【0009】次にMPEG1符号化、復号システムを図1について説明する。ビデオ入力は、前処理を経てビデオ符号化器に入り、ついでシステム多重化を経て、蓄積メディアに入りついでシステム多重分離した後、ビデオ復号器に入り前処理してビデオ出力となる。

【0010】前記MPEG1ビデオ符号化器は図2の入力画像がデータに処理される。またMPEG1ビデオ復号器は、図3のように入力バッファが表示バッファに処理される。

【0011】前述したように、MPEG1はCD-R0

Mなどの蓄積メディアに用いることが目的である。蓄積メディアでは、早送り、巻戻し、途中からの再生、逆転再生などのトリックモードが必要とされる。このようなトリックモードを実現するため、MPEG1ではグループオブピクチャー（Group of Pictures、以下GOPという）構造が取られている。

【0012】MPEG1では、符号化された画像データは、前後の画面データをもとにして作られているために、1画面だけで完結した情報にはならない。このために、何枚かの画面データをひとまとまりにしたGOPを単位として、ランダムアクセスを可能にしている。つまり、GOPの中に少なくとも1枚は、前後画面の情報を利用せず1枚だけで閉じた画面データ（Iピクチャ）を必ず含むようにすることで、このデータを元にGOP内の他の画面データの再生が可能となる。なお、1つのGOPは、通常15枚程度のピクチャをグループ化することが多い（図4）。

【0013】MPEG1では、過去再生画像からの順方向予測と未来再生画像からの逆方向予測の両方を行っている。これを双方向予測という。

【0014】双方向予測を実現するため、MPEG1では、Iピクチャ、Pピクチャ、Bピクチャの3つのタイプの画像を規定している。

【0015】これらの他に、Dピクチャ（DC符号化画像）が規定されている。これは、フレーム内の情報のみで符号化され、DCT係数の内のDC成分のみで構成されており、他の3種類のピクチャタイプと同じシーケンスに共存することはない。

【0016】MPEG1では、双方向予測を行うBピクチャが導入されることによって、予測効率が大きく向上し、高圧縮時の画質向上に役立っている。

【0017】画像データは、図6に示すように、シーケンス、GOP、ピクチャ、スライス、マクロブロック（MB）、ブロックの6層の階層構造から成っている。

【0018】前記シーケンス層とは、一続きの映像を表現するビットストリームは、シーケンスヘッダで始まり、その後1個または数個のGOPが続き、最後に1個のシーケンスエンドコードで終了する。どのGOPの直前にもシーケンスヘッダを置くことができるが、一続きの映像中のシーケンスヘッダでは量子化マトリクス以外のデータ要素は全て最初のシーケンスヘッダと同じである必要がある。

【0019】これによってシーケンス途中へのランダムアクセスが可能になる。

【0020】またGOP層とはGOPをひとつ含む。

【0021】次にピクチャ層とはIピクチャ、Pピクチャ、Bピクチャ、Dピクチャのいずれかを1枚含む。

【0022】またスライス層とは、スライスは、画像の左上から始まってラスタスキャン順に右下に続く一連の任意個のマクロブロックの集まりである。スライス間

は重なりやすき間を持たせることはできないが、スライスの位置は画面ごとで異なってもよい。スライスのデータの先頭には同期信号が割り当てられるため、復号時にデータの読みだし誤差があっても次のスライスで同期を回復できる利点がある。またスライスのデータの復号はそのスライスだけ独立して行えるため、復号の高速化のためにスライス単位に並列処理が可能である。

【0023】次にマクロブロック層とは、マクロブロックは16画素×16ラインの輝度成分と、画像中で空間位置が対応する8画素×8ラインの2つの色差成分で構成されている。ひとつのマクロブロックは4個の輝度ブロックと2個の色差ブロックからなる。マクロブロック中でのブロック順序と配置は図6のとおりである。このマクロブロックを単位に動き補償およびフレーム間予測は行われる。

【0024】更にブロック層とは、8画素×8ラインからなる輝度成分または色差成分で構成されるDCT処理単位である。

【0025】

【発明の実施の形態】この発明は、符号化された動画像を各ショットに分割し、ショット毎の類似度を用いてショットを統合し、シーンを抽出するようにした動画像の自動階層構造化の方法である。

【0026】また前記階層構造化されたデータを用いて動画像の全体的内容を把握し、また所望のシーン、ショット又はフレームの検索を容易にした動画像のブラウジング方法である。

【0027】前記この発明により、動画像の内容を把握したり、所望の場面を検索することが極めて容易となった。

【0028】

【実施例】この発明の実施例を図面に基づいて説明する。

【0029】まず、物理的な特徴量によって抽出が可能であり、検出が比較的容易な、ショットへと動画像を分割する。そして、分割されたショット間の類似度によって、ショットを統合することでシーンの抽出を行う。この際、ショットという動画像のままでは扱いにくいいため、ショット中から代表フレームをいくつか選び出す（図7）。

【0030】一般に、動画像はデータ量が多いため、その処理量は膨大なものとなる。さらに、符号化された動画像の場合、復号化が必要となるため、さらに処理量は増大する。

【0031】この発明では、MPEG1動画像を完全に復号化することなく、必要最小限の情報のみを復号化することで、高速な処理を可能としているので、フレーム間予測やDC成分の復号化による簡略画像の取得といったMPEG1の符号化アルゴリズムの特性を利用して処理量を軽減している。そこでMPEG動画像中のIピク

チャの簡略復号化とフレームの比較の方法について述べ、その後流れに沿って各処理の詳細を説明する。

【0032】動画画は多数の静止画像（フレーム）によって構成されている。したがって、動画画の解析には各フレームの画像情報は必要不可欠である。しかし、MPEG1 動画画の復号化は、比較的処理量が多く、高速な処理を実現することは難しい。

【0033】そこで、すべてのフレームではなくIピクチャだけを復号化し、さらに、完全な復号化ではなく、簡易復号化によって原フレームの縮小画像を得る。この簡易復号化は、DCT係数のDC成分を復号化すると元のブロックの平均色が得られることを利用する。つまり、各ブロックのDCT係数のうちDC成分だけを復号化し、得られた平均色で各ブロックを代表させた画像を作るのである（図8）。

【0034】このようにして得られた画像をDC画像と呼ぶこととする。各ブロックの大きさは 8×8 であるから、DC画像は縦横それぞれ原画像の $1/8$ の大きさとなる。

【0035】BピクチャおよびPピクチャの復号には、自分自身だけでなく、動きベクトル情報および参照先のピクチャなど直接使わない情報の復号が必要となるが、Iピクチャはフレーム内で閉じた符号化がなされているので、そうした情報の復号化の必要はない。また、Iピクチャ中のイントラ（Intra）マクロブロックのDC成分は、計算量の多いIDCTを行うことなく下式（1）によって復号化することができる。したがって、DC画像は非常に高速に得ることができる。

【0036】

【数1】

$$\begin{aligned} Y_k &= (DY_k)/8 \\ Cb_{k'} &= (DCb_{k'})/8 \\ Cr_{k'} &= (DCr_{k'})/8 \end{aligned} \quad (1)$$

【0037】ここで、 Y_k 、 $Cb_{k'}$ 、 $Cr_{k'}$ は各ブ

$$D_{histarea} = 1 - \sum_{n=0}^{N_{bin}} \min(I_n, J_n) / \sum_i^{N_{bin}} I_n \quad (2)$$

【0044】

$$D_{pixsum} = \frac{1}{H \cdot W} \sum_{i=0}^H \sum_{j=0}^W |a_{ij} - b_{ij}| \quad (3)$$

【0045】ここで、ふたつの値から類似度を算出する手法として、簡略化ファジィ推論を用いる。ファジィ推論を用いる事で、色ヒストグラムによる距離および画素値による距離と、画像の類似度の関係を厳密に定式化する事なく記述できる。また、簡略化ファジィ推論による

ロック（ k 、 k' はブロック番号）の平均色の輝度および色差成分、 DY_k 、 $DCb_{k'}$ 、 $DCr_{k'}$ は各ブロックのDC成分である。

【0038】実際に、約30分のMPEG1 動画画（GOPは図5のタイプのもの）について、全てのフレームを復号した場合と、IピクチャのDC画像だけを復号した場合の処理時間を表1に示す。全てのフレームを復号する場合に比べ、約 $1/20$ の処理時間で復号できることがわかる。

【0039】

【表1】

表1 復号時間の比較

画像の長さ	1832.0
全てのフレームを復号	621.7
Iピクチャだけを復号	154.3
DC画像だけを復号	32.6

【0040】また、図9に、DC画像の例と、その原画像を示す。

【0041】フレーム間の比較に用いる類似度は、画素値の比較と色ヒストグラムの比較に大別される。色ヒストグラムによる比較はカメラや被写体の動きに影響を受けにくいために、類似度として用いるには都合が良いが、半面、空間的な情報を全く含まないために全く違う画像が同じ色ヒストグラムを持つ場合が問題となる。色ヒストグラムによる比較に空間的な情報を持たせようとする試みはいくつかなされているが、いずれもある程度複雑な処理を必要とする。

【0042】この発明では、色ヒストグラムによる距離として式（2）の $D_{histarea}$ を、画素値による距離として式（3）の D_{pixsum} を用い、このふたつの値を組み合わせることで類似度を算出することで処理の単純さを損なわずに空間的な情報を加味した類似度を求める。

【0043】

【数2】

【数3】

推論は単純で、高速に実行できる。

【0046】このとき用いるファジィルールは以下式（4）の通りである。

【0047】

【数4】

$$\begin{array}{llll}
 \text{rule } i: & \text{IF} & D_{histurea} & \text{is } A_a \\
 & & D_{pizsum} & \text{is } B_b \\
 & \text{THEN} & & s_i = c_i \\
 & & a, b = \text{small, medium, large} \\
 & & i = 1, 2, \dots, 9
 \end{array} \quad (4)$$

【0048】ここで、 i はルール番号、 I はルール数、 c_i は後件部を表す実数値であり、 $[0, 1]$ の値をとる。また、 A_a 、 B_b はそれぞれその特徴値のメンバシップ関数であり、各特徴値ごとに図10のような“small”、“medium”、“large”の3つのメンバシップ関数を設定する。

$$w_i = \mu_{A_a}(D_{histurea}) \cdot \mu_{B_b}(D_{pizsum}) \quad (5)$$

【0051】

【数6】

$$s = \frac{\sum_{i=1}^I w_i c_i}{\sum_{i=1}^I w_i} \quad (6)$$

【0052】ショットの検出はショットの間のカット点の検出を行う。

【0053】カット点の検出とは、フレーム間の相関の低い点を検出する作業に他ならないが、この発明では、フレームの特徴量を直接比較して相関を調べるのではなく、MPEG1の符号化の様子から相関を調べ、カット点を検出する。つまり、MPEG1において、フレーム間の相関から予測によって圧縮が行われていることを利用し、逆に、予測の行われ方を調べることでフレーム間の相関を調べるのである。

【0054】新たにフレームの特徴量を調べることなく、MPEG1の符号化情報を利用することで、計算量が少なくて済み、また、すべての情報を復号化する必要がないため、高速な処理が可能である。

【0055】処理手順としては、まず、Bピクチャにおける参照の様子からカット点を検出し、さらにPピクチャでの参照、Iピクチャの変化を調べて確認を行う。図11のような $N=15$ 、 $M=3$ のGOPを例に説明する。

【0056】Bピクチャでは、前後両方のIまたはPピクチャから参照を行なっている。即ちBピクチャ中には、一般的に(I MB)、(F MB)、(B MB)、(Bi MB)の4種類のマクロブロックMBが存在し、それぞれ、参照を全く行わないI MBか、順方向F MB、逆方向B MB、双方向I n MBの参照を行っている。このときの参照の様子は図12のようになる。

【0057】ショットの中、すなわちフレーム間の相関が高い場合には、過去および未来への参照の数はほぼ等しいが、参照するフレームとの間にカット点が存在すると、過去または未来へ依存が大きく偏り、マクロブロックの構成に偏りが生じる。このときの様子を図12

【0049】このルールに対する適合度を式(5)により求め、次に式(6)で最終的な推論結果、すなわち画像間の類似度 s を求める。なお、 s は $[0, 1]$ の値を取る。

【0050】

【数5】

(b)(c)(d)に示す。ただし、図12は極端な場合であり、実際にはカット点を越えた参照が完全になくなる訳ではない。

【0058】このことからわかるように、Bピクチャのマクロブロックタイプの構成から、Bピクチャの前後フレームへの参照の様子を判断することができる。これをBピクチャの依存度 $relat$ として式(7)のように定義する。

【0059】

【数7】

$$relat = \frac{N_F - N_B}{N_{Bi}} \quad (7)$$

【0060】ただし、 N_F 、 N_B 、 N_{Bi} はそれぞれBピクチャに含まれるF MB、B MB、Bi MBの数である。

【0061】 $relat$ は、 N_F と N_B の差が大きく、また N_{Bi} が少ない程、その絶対値が大きくなり、参照の偏りが大きいことを示す。

【0062】さて、図11のようなGOPにおいて、ふたつのPピクチャ(またはIピクチャ)とそれに挟まれたふたつのBピクチャに注目し(例えば f_7 、 f_8 、 f_9 、 f_{10})、これを $P_1 B_2 B_3 P_4$ と表すことにすると、全てのカット点は必ず $P_1 | B_2 B_3 P_4$ 、 $P_1 B_2 | B_3 P_4$ 、 $P_1 B_2 B_3 | P_4$ のいずれかの形で現れる($|$ はカット点を表す)。これらはそれぞれ図12の(b)(c)(d)に対応する。このとき、どの場合でも、図12からわかるとおり、 B_1 、 B_2 の両方に参照の偏りが生じ、依存度の絶対値が大きくなる。

【0063】そこで、次式(8)を満たせば、 $P_1 | B_2 B_3 P_4$ 、 $P_1 B_2 | B_3 P_4$ 、 $P_1 B_2 B_3 | P_4$

のいずれかの形でカット点が存在すると判断する。

【数8】

【0064】

$$|relat_{B_1} \cdot relat_{B_2}| > threshold_B \quad (8)$$

【0065】次に、 $P_1 \mid B_2 \ B_3 \ P_4$ 、 $P_1 \ B_2 \mid B_3 \ P_4$ 、 $P_1 \ B_2 \ B_3 \mid P_4$ のどのパターンかを判断し、正確なカット点を決定する。式7からもわかるとおり、 $relat$ は、過去からの参照が多いと正、未来か

らの参照が多いと負の値を取る。これを利用して、式(9)のようにカット点を決めることができる。

【0066】

【数9】

$$\begin{aligned} P_1 \mid B_2 B_3 P_4 & \text{ if } relat_{B_1} < 0, & relat_{B_1} < 0 \\ P_1 B_2 \mid B_3 P_4 & \text{ if } relat_{B_1} > 0, & relat_{B_1} < 0 \\ P_1 B_2 B_3 \mid P_4 & \text{ if } relat_{B_1} > 0, & relat_{B_1} > 0 \end{aligned} \quad (9)$$

【0067】以上のようにして、Bピクチャの参照情報からカット点を検出することができる。

【0068】Bピクチャの参照による検出だけでは、ノイズや、カメラの前を物体が横切るなど瞬間的な画面の変動がある際に、誤検出が発生することがある。これは、Bピクチャと参照先のピクチャとの距離が短いと考えられる。そこで、Bピクチャだけではなく、より遠いピクチャを参照するPピクチャの参照情報を利用して結果の確認を行う。

【0069】Pピクチャは、参照しているIまたはPピクチャとの間にカット点が存在すると、参照がほとんどできなくなるため、IMBの数が増加するはずである。そこで、次式(10)を満たす場合は、間にあるBピクチャから求めたカット点は誤検出であるとみなし、これを除去する。

【0070】

【数10】

$$\frac{N_f}{N} < threshold_P \quad (10)$$

$$D_{histarea} < threshold_I$$

【0076】次にショットの代表フレームの選出について説明する。ここに代表フレームとはショットは動画全体に比べれば短い単位ではあるが、例えば5秒間のショットでは150枚(30fpsの場合)のフレームの集合であり、このままでは、比較、表示、特徴値の検出などの処理がしにくい。そこで、一般に、ショットを扱う際には、ショットの中からそのショットを代表するフレームを選び出し、この代表フレームによって比較、表示などの処理を行う。

【0077】この発明においては、ショットを統合しシーンを抽出する際に、ショット間の類似度を求めるために用いる。また、解析された動画構造をユーザに提示する際にショットの内容を簡単に示すために用いられる。したがって、ショットの内容を最もよく表しているフレームを選ぶことが必要となる。

【0078】ショットを扱っている従来の研究では、この代表フレームとして、機械的にショットの先頭のフレ

ームただし、 N_I はPピクチャに含まれるIMBの数、 N はPピクチャ中の全MBの数である。

【0072】Pピクチャよりさらに離れたIピクチャどうしの比較を使った確認を行う。

【0073】また、図11のタイプのGOPの場合、 f_3 のIピクチャの前にある2つのBピクチャ(f_1 、 f_2)によるカット点に対しては、Pピクチャを利用する結果の確認は行うことができない。 f_3 のIピクチャは参照を行わないからである。この部分で起こる誤検出の検出のためにもこのIピクチャによる結果の確認が必要となる。

【0074】Iピクチャ(f_3)とひとつ前のGOPにおけるIピクチャ、それぞれのDC画像に対する色ヒストグラム距離 $D_{histarea}$ (式(2))を調べ、次式(11)を満たす場合は間にあるBピクチャから求めたカット点は誤検出であるとみなし、これを除去する。

【0075】

【数11】

$$(11)$$

ームあるいは中央のフレームを用いているものが多い。しかし、そのようにして選ばれたフレームはショットの内容をよく表すとは言い難い。そこでこの発明では、ショットに含まれるフレームの平均に最も近いフレームを代表フレームとして選ぶこととする。

【0079】また、ショットはほぼ動きがない場合だけではなく、

(1) ひとつのショットの途中でカメラの動き(パン・ズームなど)があるもの。

(2) カメラあるいは画像中のオブジェクトが動き続けているもの。

(3) 動きが非常に激しいもの。

などの場合がある。このようなショットではひとつのフレームでショット全体を代表させるのは難しく、有用な情報を落とす危険がある。そこで、このようなショットは、複数の代表フレームによって表すこととする。

【0080】さらに、複数の代表フレームを選出するこ

とによって、カット点の検出ミスによる影響を少なくすることができる。つまり、カット点の検出ミスにより、本来複数であるショットがひとつにまとまってしまった場合、代表フレームをひとつだけ選出すると、本来、ただひとつのショットの情報だけしか使われないことになる。これに対し、内容に基づいて複数選出すれば、それぞれのショットの情報を捨てることなく、シーン抽出の際に生かせることになるわけである。

【0081】必要最小限の代表フレームを選び出すため、まず、ショット中のフレームのクラスタリングを行う。この結果できた各クラスタからそれぞれもっとも平均に近いフレームを選び出し、これをショットの代表フレームとする。

【0082】ただし、ショット内のすべてのフレームを代表フレームの候補とすると、ショットが長くなったときに処理量が増大する恐れがある。そこで、候補としてIピクチャだけを用いる。これにより、選出のための処理量だけでなく、原動画像からの復号化のための処理量も削減することができる。また、MPEG1において一般に、符号化効率を上げるためにI、P、Bピクチャの

$$N_I < threshold_{move}$$

【0088】(3) 前記初期クラスタをもとにクラスタリングを行い、ショット中のIピクチャをいくつかのクラスタへと分類する。クラスタリングは群平均法を使って行い、要素間の距離としては $D_{histarea}$ (式(2))を用いる。クラスタリングは、クラスタ間の距離のうち最小のものが閾値を越えるまで行う。

$$\begin{aligned}\overline{Y}_{ij} &= \frac{1}{N} \sum_{n=1}^N Y_{ij}^n \\ \overline{Cb}_{ij} &= \frac{1}{N} \sum_{n=1}^N Cb_{ij}^n \\ \overline{Cr}_{ij} &= \frac{1}{N} \sum_{n=1}^N Cr_{ij}^n\end{aligned}\quad (13)$$

ここで、 N はDC画像の数、 \overline{Y}_{ij} 、 \overline{Cb}_{ij} 、 \overline{Cr}_{ij} は平均DC画像の座標 (i, j) における輝度及び色差信号、 Y_{ij}^n 、 Cb_{ij}^n 、 Cr_{ij}^n は n 番目のDC画像 DC_n の座標 (i, j) における輝度及び色差信号を表す。

【0091】(5) 平均DC画像との距離 D_{pixsum} (式(3))が一番小さいDC画像 DC_k をもつIピクチャ I_k を代表フレームとする。

【0092】以上のようにしてショットの代表フレームが選出される。

【0093】実際の処理では、IピクチャのDC画像や、PおよびBピクチャのマクロブロック情報を効率的に得るために、代表フレームの選出はカット点の検出と並行して行われる。

【0094】たとえば会話のシーンなどでは、話者を交

量子化特性を変えることが多いため、Iピクチャが最も品質がよい場合が多いことも都合がよい。

【0083】さらに、Iピクチャを完全に復号化するのではなく、前記で述べたDC画像を用い、復号時の処理量削減を計る。

【0084】具体的な処理手順は以下になる(図13)。なお、Iピクチャがひとつも含まれないショットの場合、つまり非常に短いショットの場合はショット中で一番初めに現れるPピクチャを、それも存在しない場合はBピクチャを代表フレームとする。非常に短いショットの場合、ショット中での変化はほとんどないといえるから、このような機械的な処理で十分である。

【0085】(1) ショットに含まれるIピクチャを簡易復号化し、DC画像を取り出す。

【0086】(2) ショット中、動きが少ない部分のDC画像を初期クラスタとする。動きが少ないかどうかは、BおよびPピクチャに含まれるIMBの数を調べることによって行う(式(12))。

【0087】

【数12】

(12)

【0089】(4) クラスタリング終了後、各クラスタからそれぞれひとつずつ代表フレームを選び出す。まず、クラスタ内のIピクチャのDC画像を平均して、平均DC画像をつくる(式(13))。

【0090】

【数13】

互に撮る場合が多いため、同じようなショットの繰り返しになる。このように、ひとつのシーンのなかには、似ているショットがいくつか含まれることが多い。この性質に着目してショットを統合し、シーンを抽出する。

【0095】ショット間の類似度は、それぞれのショットの代表フレーム間の類似度 s (式(6))を用いる。ただし、ひとつのショットが複数の代表フレームを持つ場合もあるため、すべての代表フレームの組み合わせについての類似度を調べ、そのうちの最大値をショットの類似度とする(図14)。

【0096】ショット間の類似度からシーンを抽出する最も簡単な方法は、図15のように、似ている（類似度が非情に高い）ショットが存在すれば、その間をすべて同じシーンとみなす方法である。

【0097】しかし、このようにすれば、

- (1) 似ているか、似ていないかの閾値の設定が結果に大きく影響する
- (2) 類似度が非常に高いショットの組はないが、中程度の類似度のショットの組が多数ある、といった場合で

も同じシーンとみなすことができず、柔軟性に欠けるといった問題点がある。

【0098】そこで、ショット $shot_n$ とショット $shot_{n+1}$ が連続している（すなわち、同じシーンに属する）度合を表す結合度 $connect_{n, n+1}$ を式(14)のように定義し、この結合度を用いてシーンの抽出を行う。

【0099】

【数14】

$$connect_{n,n+1} = 1 - \prod_{i=n-N+1}^n \prod_{j=n+1}^{i+N} (1 - s_{ij}) \quad (14)$$

【0100】ここで、Nは比較するショットの範囲を表す。 s_{ij} は $shot_i$ と $shot_j$ の類似度である。

【0101】このように、結合度 $connect_{n, n+1}$ はショット $shot_n$ とショット $shot_{n+1}$ だけでなく、その付近のすべてのショット間の類似度 s_{ij} から求められる。例えば、図16において、 $connect_{3, 4}$ はショット $shot_3$ とショット $shot_4$ の類似度 s_{34} だけでなく、ショット $shot_2$ とショット $shot_5$ の類似度 s_{25} も使って求められる。なぜなら、たとえショット $shot_3$ とショット $shot_4$ がまったく類似していなくても、ショット $shot_2$ とショット $shot_5$ が類似していれば、ショット $shot_3$ とショット $shot_4$ は同じシーンに属すると考えられるからである。

【0102】ただし、時間的に遠く離れているショット同士は、同じシーンに属する可能性が小さく、むしろ違うシーンに属するにもかかわらずたまたま類似度が高いショットが存在する可能性があり、このような原因による未検出をできるだけ防ぐため、比較するショットの範囲はNに制限する。

【0103】式(13)によって得られる結合度 $connect_{n, n+1}$ の変化は、例えば図17のようになる。

【0104】このような結合度の変化から、シーンチェンジを決定しシーンを抽出する。ここでは、変化のピークと谷の差が閾値 $threshold_{SCENE}$ より大きいとき、その谷となる結合度をもつカット点をシーンチェンジ点とする。

【0105】

【発明の効果】この発明によれば、与えられた動画像をハイブリッド符号化し、これを階層的構造へ分割し、分割されたショット間の類似度によりショットを統合してシーンを抽出するので、シーンの抽出が、迅速、正確に行われる効果がある。然してシーンからショット又はフレームを抽出するのは比較的容易であるから、結局動画像からシーン、ショット又はカットを短時間に、かつ正確に抽出し得る効果がある。

【0106】前記処理は総て現在使用されているハードに、適切なソフトを組み込むことにより自動化できるので、適切な入力指示により、所望のシーン、ショット又はカットを自動的に提供できる効果がある。

【0107】実験の結果によれば、表2の動画像を用いたカットの検出結果は表3の通りである。

【0108】

【表2】

表2 評価に用いた動画像

	題 名	長さ(分)	フレーム数	画面サイズ	ソース
A	A Room with a View	31:52	45838	352×240	VideoCD
B	Kramer v.s. Kramer	30:09	43379	352×240	VideoCD
C	Stand by Me	30:32	43942	352×240	VideoCD

【0109】

【表3】

表3 カット点検出の結果

動画像	カット点の数	検出数	未検出数	誤検出数	検出率(%)	誤検出率(%)
A	272	270	2	11	99.2	4.04
B	267	253	14	0	94.7	0
C	377	259	18	0	95.2	0

【0110】またカット点検出の処理時間は表4の通り

である。

【0111】

【表4】

表4 カット点検出の処理時間

動画像	処理時間(sec.)
A	110.44
B	107.09
C	98.74

【0112】更に単純なアルゴリズムによるカット点検出の処理時間は表5の通りである。

【0113】

【表5】

表5 単純なアルゴリズムによるカット点検出の処理時間

動画像	処理時間(sec.)
A	5160
B	5441
C	5161

表6 シーンチェンジ点検出の結果

動画像	シーンチェンジ点の数	検出数	未検出数	誤検出数
A	17	17(4)	4	7
B	22	21(5)	1	11
C	24	18(6)	6	3

【0118】構造解析処理全体の処理時間は表7のようになり、十分な高速性を保っていることがわかる。

【0119】

【表7】

表7 構造解析処理の処理時間

動画像	処理時間(sec.)
A	128.2
B	127.1
C	126.5

【図面の簡単な説明】

【図1】この発明のMPEG1符号化・復号システム。

【図2】同じくMPEG1ビデオ符号化器のブロック図。

【図3】同じくMPEG1ビデオ復号器の図。

【図4】同じくGOPの例示図。

【図5】同じく原画像およびストリーム上の画面の並びを示す図。

【図6】同じくMPEG1の階層構成図。

【図7】同じく構造解析処理の流れ図。

【図8】同じくDC画像の生成図。

【図9】同じくDC画像の例示図。

【図10】同じく類似度を求めるファジィ推論に用いるメンバシップ関数の形状図。

【0114】次にカット点検出の結果得られたショットを用いて、シーン抽出を行った。結合度の変化を図18、19、20に示す。

【0115】図18、19、20の結合度からシーン抽出を行った結果が表6である。なお、シーンチェンジ点検出の閾値は、 $\text{threshold}_{\text{SCENE}} = 0.3$ とした。

【0116】表6から、いずれの動画像についてもシーンチェンジ点のうち75%以上を検出できており、高速性や本手法が意味解析や知識を使っていないことを考慮すると十分実用的であるといえる。なお、検出数の内1/4から1/3程度は、実際のシーンチェンジ点から1ショット分前または後ろにずれて検出されている。これは、シーンチェンジ点に隣接するショットに関して、その本来属すべきシーンのなかにそのショットへの類似度が高いショットが存在しない場合は結合度が低くなってしまうというアルゴリズム上の欠点による。

【0117】

【表6】

【図11】同じくGOPの例示図。
 【図12】(a)同じく通常の参照図。
 (b)同じく過去のPピクチャとの間にカット点がある場合を示す図。
 (c)Bピクチャの間にカット点がある場合を示す図。
 (d)同じく未来のPピクチャとの間にカット点がある場合の図。

【図13】同じく代表フレームの選出アルゴリズムを示す流れ図であって、
 (a)DC画像を取り出す図。
 (b)初期クラスタを決定する図。
 (c)クラスタリングの図。
 (d)クラスタ中のDC画像から平均画像を作る図。
 (e)平均画像に最も近いものを代表フレームとする図。

【図14】同じくショット間の類似度を示す図。

【図15】同じく単純なシーン抽出の図。

【図16】同じく $N=3$ のときの結合度を示す図。

【図17】同じく結合度 $\text{connect}_{n, n+1}$ の変化の例示図。

【図18】同じく動画像Aの結合度の変化を示す例示図。

【図19】同じく動画像Bの結合度の変化を示す例示図。

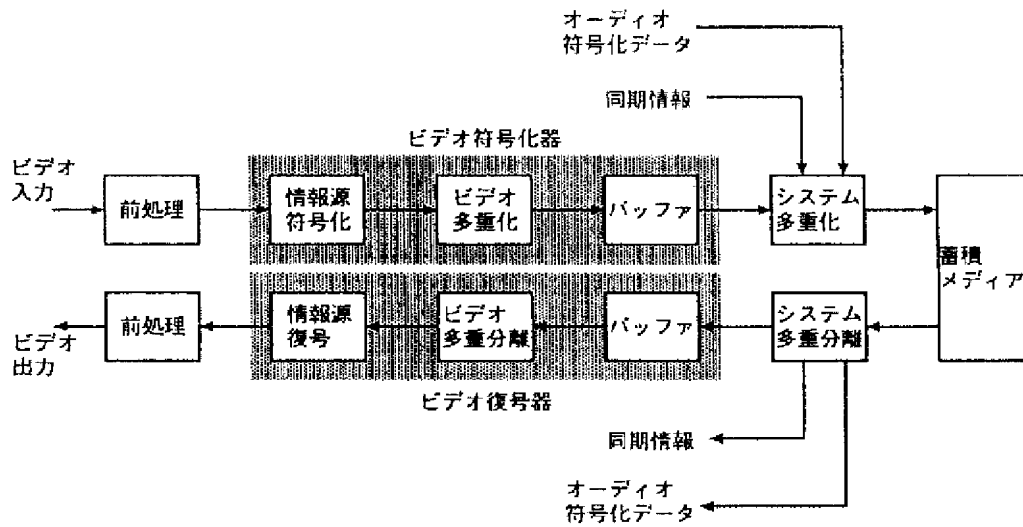
【図20】同じく動画像Cの結合度の変化を示す例示図。

【図18】同じく動画像Aの結合度の変化を示す例示図。

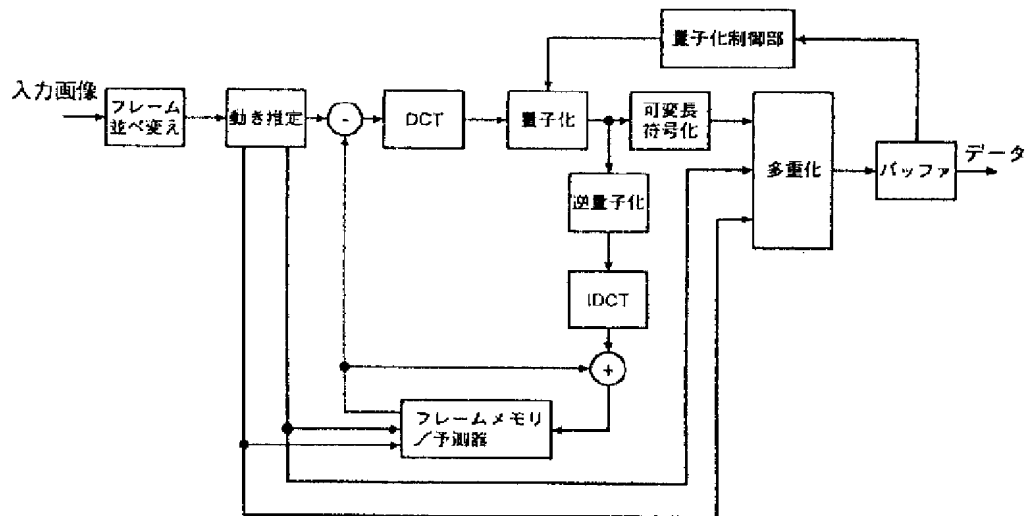
【図19】同じく動画像Bの結合度の変化を示す例示図。

【図20】同じく動画像Cの結合度の変化を示す例示図。

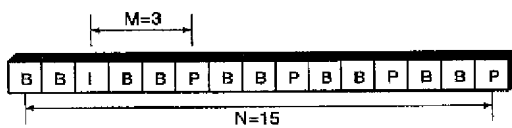
【図1】



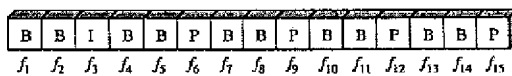
【図2】



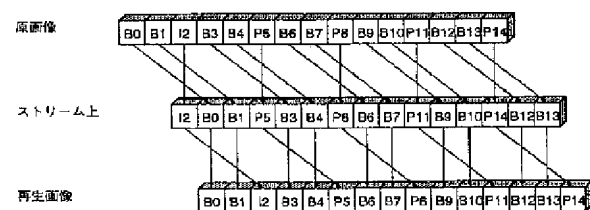
【図4】



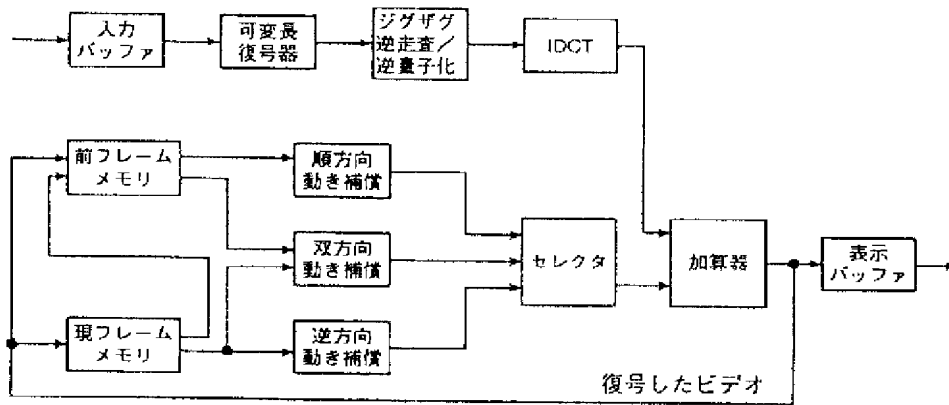
【図11】



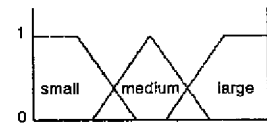
【図5】



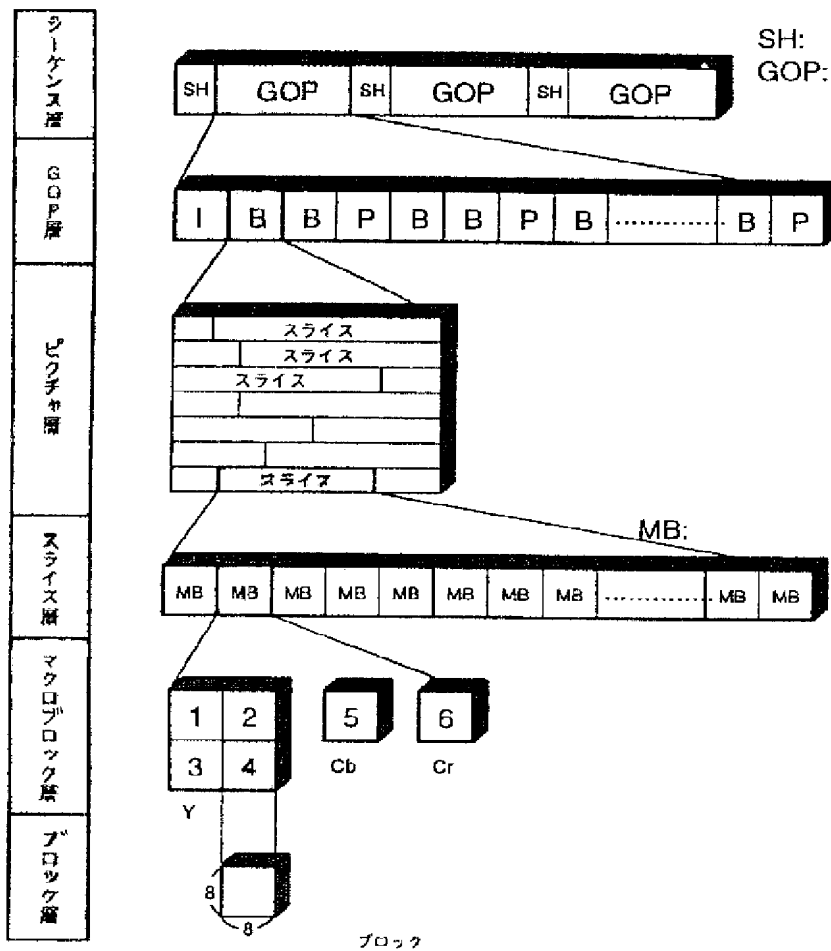
【図3】



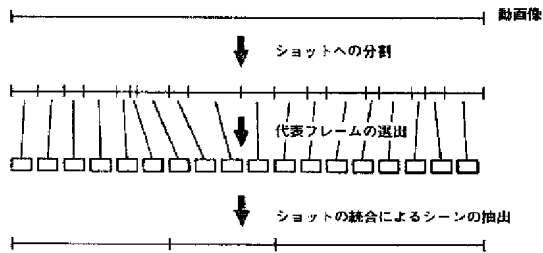
【図10】



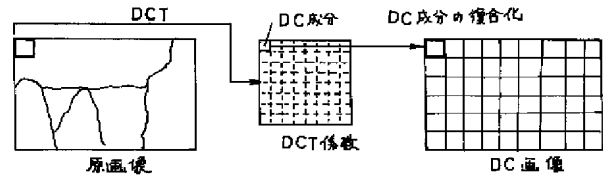
【図6】



【図7】

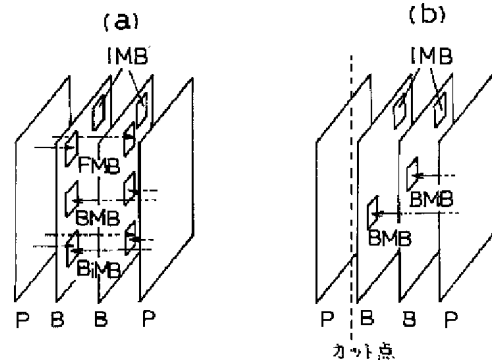
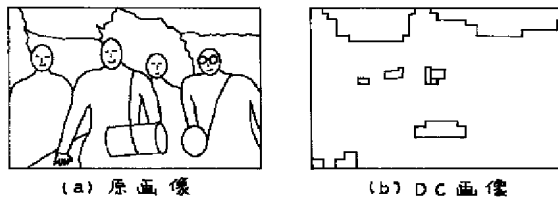


【図8】

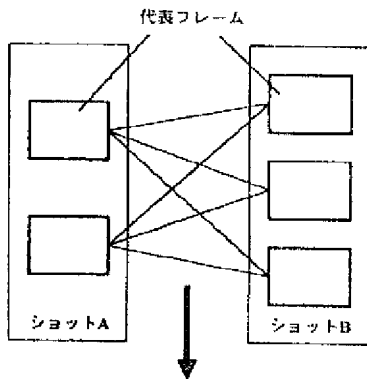


【図12】

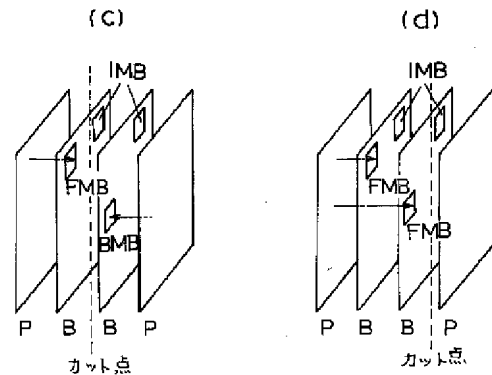
【図9】



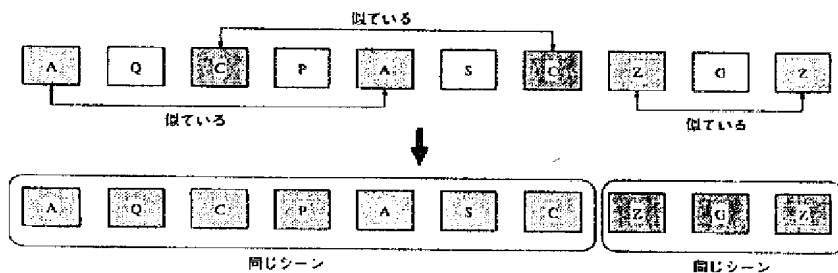
【図14】



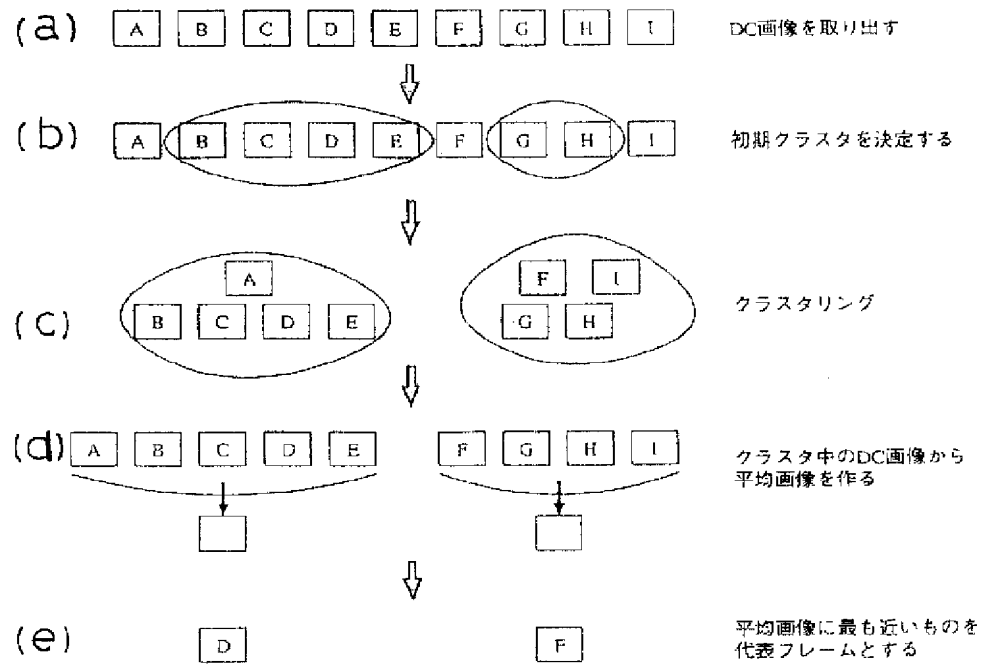
全ての組合せの代表フレーム間の類似度のうち
最大値をショット間の類似度とする



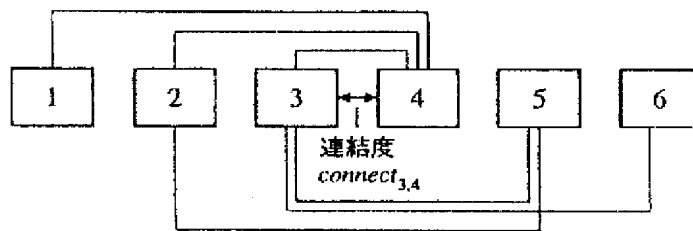
【図15】



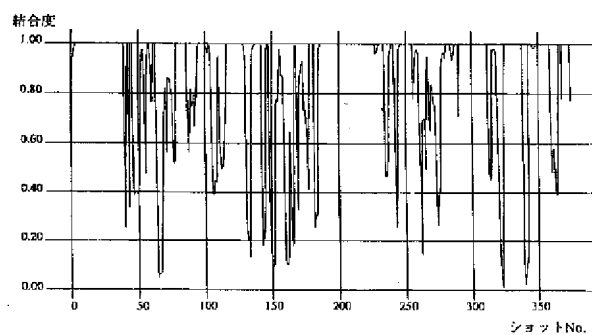
【図13】



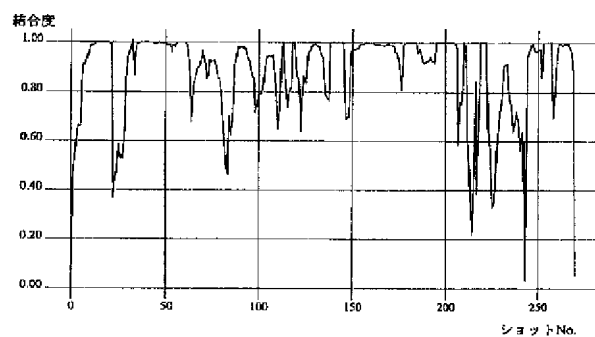
【図16】



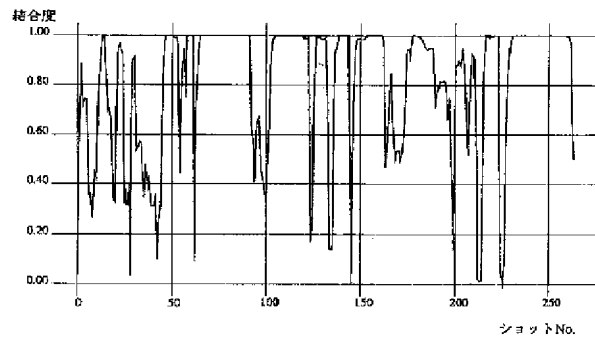
【図17】



【図18】



【図19】



【図20】

